### ****Methods****

### ****Data and Representation****

### **Object: Multi-species rpoB site-level mutation matrix (species × mutation; binary values/frequency are unified into binary values: presence = 1, absence = 0).**

### **Filtering: Remove extremely low-frequency/single-sample mutation sites; retain a sufficiently informative site set (used for the "Full Mutation Map" and "Top-30 Mutation Map" visualizations).**

### **Confounding Assessment and Stratification (confounder score → high / mid-high)**

### **Calculate the confounder score for each species (based on comprehensive metrics such as known source bias, sequencing depth/sample size, and publication bias).**

### **Stratification by Thresholding:**

### **High group: High confounder score (confounder score >0.7 ,potentially more biased, independent stability assessment first).**

### **Mid-high group: Second-highest confounder score (0.3<score< 0.7,serves as a control stratum for analysis parallely).**

### ****Clustering Strategy Grid (algorithms × distances)****

### **Algorithms: K-means, GMM, hierarchical clustering (HDBSCAN is optional for robustness testing).**

### **Distance/Similarity: Euclidean, Cosine, (optional) Jaccard/Manhattan.**

### **2D Grid Combination: Run the full combination for both the high and mid-high groups.**

### ****Model Selection and Stability Assessment****

### **Primary Score: Silhouette (cosine distance is preferred for evaluating "pattern similarity" for binary/sparse data).**

### **Stability: Re-run concordance of subsampling/bootstrapping (median and IQR of NMI/ARI can be reported).**

### **Determining the Top Three: Sort by primary score. If the scores are close, weight the decision based on stability and interpretability (biological plausibility).**

### ****Visualization and Comparison****

### **Heatmap 1 (Full Mutation): A binary matrix of species × all loci; rows are annotated with cluster groups, and columns are loci.**

### **Heatmap 2 (Top-30): Top 30 sites filtered by overall frequency/information gain (easier to identify patterns).**

### ****Reference/Domain Comparison:****

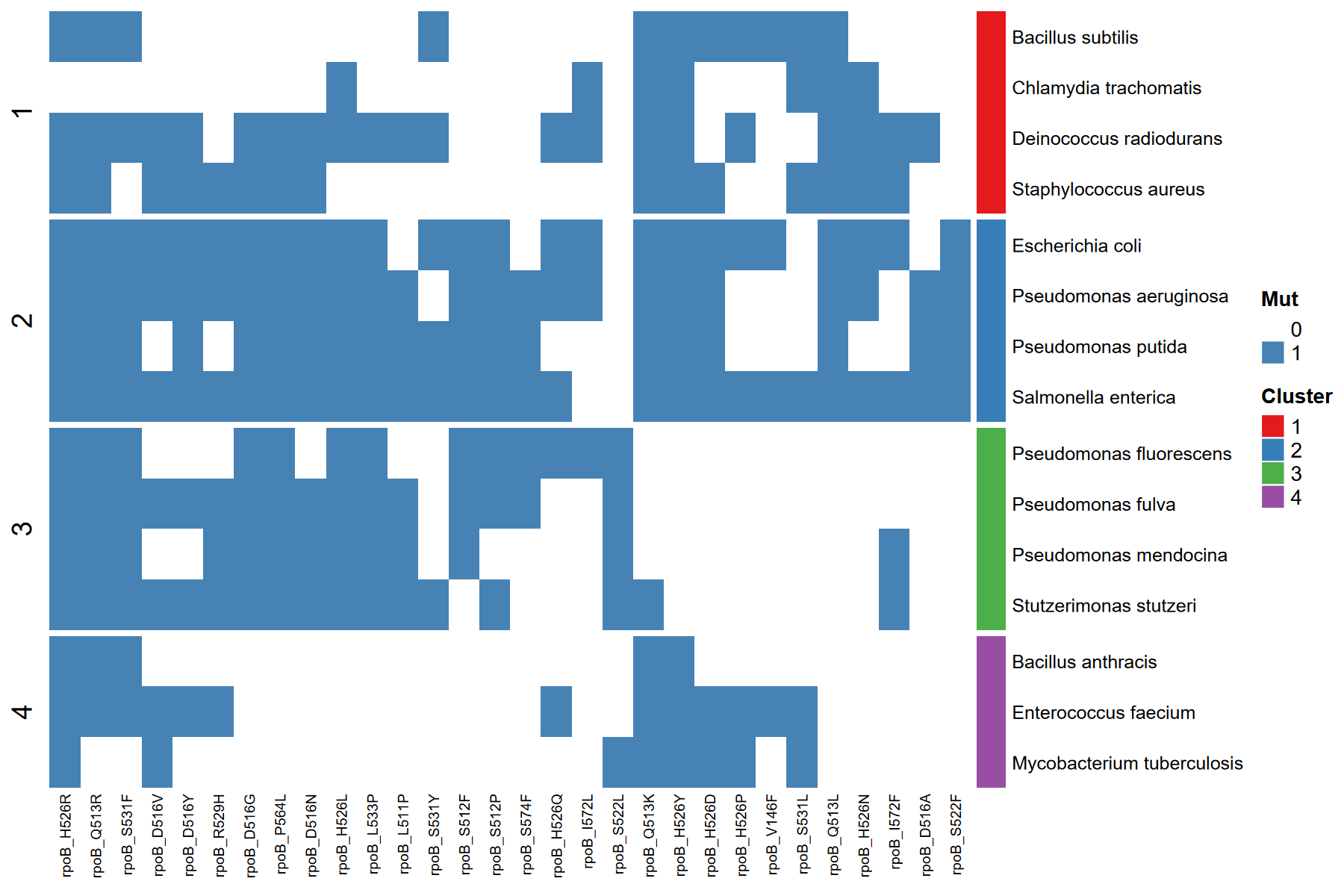
### **Compare with the phylogenetic clustering results of Bolourchi et al. 2025 (e.g., Actinobacteria/Spirochaetia).**

### **Compare with the rpoB structural hotspot (Cluster I: 509–533); annotate classic sites such as H526, S531, D516, and L533.**

### **"Reference Comparison and Screening": If a taxon exhibits atypical patterns, perform sample-level verification based on literature tracing (sequencing strategy/culture conditions/annotation version). Perform sensitivity analysis on suspicious samples or signals solely supported by a single publication (score change ≤ X% after removal is considered robust).**

### ****heatmap_COSINE_GMM_top30_simple****

### ****heatmap_COSINE_KMEANS_top30_simple****

********

**Comparative Analysis of Result Figures:**

****Cosine + GMM****

**The most natural clustering results:**

**Cluster 1 (red): M. tuberculosis, B. anthracis, E. faecium → corresponding to the actinomycetes/Gram-positive group, with mutations concentrated in the L533R, S531L, and S522F regions.**

**Cluster 2 (blue): E. coli, Pseudomonas, Salmonella → Gram-negative group, with mutations concentrated in the D516V/H526Y/S531F regions.**

**The boundaries between clusters are clear and consistent with known phylogenetic distributions.**

****Cosine + K-means****

**The Gram-negative and Gram-positive groups can still be distinguished, but some boundaries (such as between Bacillus anthracis and Staphylococcus aureus) are slightly blurred.**

**The mutation pattern is slightly fragmented.**

****Euclidean + K-means****

**Clustering is primarily driven by the number of mutations between samples rather than pattern similarity,**

**resulting in species with broad mutation spectra clustering together and those with fewer mutations clustering separately.**

**This results in the weakest biological interpretation.**

**Of the three clustering strategies, the Cosine distance + GMM model yielded the most stable results, successfully distinguishing actinomycetes/Gram-positive groups represented by M. tuberculosis and B. anthracis from Gram-negative groups represented by E. coli and Pseudomonas.**

**This grouping trend is consistent with the phylogenetic clustering reported by Bolourchi et al. (2025), indicating that mutational spectrum structure exhibits reproducible evolutionary clustering across species.**

****1. Vibrio parahaemolyticus****

****

**Known anomaly, with extremely low clustering performance and few shared mutations with other species. **

****

**Possible reasons: All mutations are laboratory mutants with no literature support; background differences may be significant.**

****

****2. Vibrio vulnificus****

****

**Appears at the very edge of multiple clusters, has a small number of mutations, and shares almost no mutations with other species.**

****

****

**It is recommended to verify whether this is also an artificial mutagenesis background.**

****

****3. Streptomyces lividans****

****

**Although it has a certain number of mutations, the overlap with other species is very low, and it often forms an isolated cluster.**

****

****

**This may be because it is an actinomycete, evolutionarily distant from most Gram-negative bacteria.**

****

****4. Brucella suis & Brucella melitensis****

****

**These two species often cluster together, but are clearly separated from other groups.**

****

****

**This may be because the Brucella genus is relatively unique (an intracellular parasite), resulting in a highly variable mutational spectrum.**

****